# 基于用户兴趣的跨网络用户身份识别算法

邓诗琦，李　雷，施化吉

(江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013)

**摘　要：**跨网络用户身份识别的研究不但为多网络数据融合提供了依据，而且在用户身份监管、谣言控制等领域均有广泛应用价值。针对现有算法对用户兴趣在跨网络用户身份识别中作用的忽视以及时间复杂度高的问题，提出了基于用户兴趣的跨社交网络用户身份识别算法（UI-UI）。首先利用分块（blocking）思想对用户节点进行初筛选，以提升算法效率降低时间复杂度；其次根据用户产生内容（user generated content，UGC）和用户社交关系对用户兴趣进行建模，并计算兴趣相似度作为身份识别的依据；最后利用半监督学习的方法进行跨网络用户身份识别。通过在真实社交网络中进行实验，结果表明 UI-UI 算法能有效识别跨网络用户，且准确率和召回率稳定，运行时间显著减少。

**关键词：**跨网络用户身份识别；分块；用户兴趣

**中图分类号：**TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.08.0617

## User identification across social networks based on user interests

Deng Shiqi, Li Lei, Shi Huaji

(*School of Computer Science & Communication Engineering, Jiangsu University, Zhenjiang Jiangsu* 212013, *China*)

**Abstract:** The research of user identification across social networks not only provides a basis for multi-network data fusion, but also has a wide range of applications in user identity monitoring, rumor control and other fields. Aiming at the problem of ignoring the role of user interest in user identification across social networks and the high time complexity, this paper proposed a user identity algorithm based on user interest (UI-UI) . Firstly, the proposed algorithm filtered the user nodes by Blocking to improve the efficiency of the algorithm and reduce the time complexity. Secondly, it modeled the user's interest according to the user generated content(UGC) and user social relations, and used the similarity of user interest as the basis for user identification. Finally, it used the method of semi-supervised learning for user identification. Experiments on real social networks show that UI-UI algorithm can effectively identify cross-network users, and both the accuracy and recall rate of the algorithm are stable, besides, the running time is significantly reduced.

**Key words:** user identification across social networks; blocking; user interests



## 0　引言

随着互联网的迅猛发展以及智能终端的日益普及，Twitter、Facebook、新浪微博和人人网等社交网络已经成为人们信息获取与交流的主要渠道。人们会因不同社交需求同时参与多个网络，一份社交媒体研究报告指出截至 2014 年已有 52%的在线成人使用了两个或更多的社交网站。识别出这部分跨网络用户在不同网络中的账号就是跨网络用户身份识别问题，解决该问题为各类挖掘和学习任务提供了新的机遇和挑战。跨网络用户身份识别技术可以获取用户全面的社交行为模式，为用户行为深入分析以及广告精准投放提供依据；解决推荐系统中数据稀疏和冷启动的问题，为用户提供个性化推荐服务[1~3]；反映网站的发展兴衰以及帮助分析用户在网络之间的迁移模式[4]；除此之外其在商业、网络安全、信息检索等领域也有广泛的应用。

现有跨网络用户身份识别算法通常分为三类：基于用户属性的方法[5,6]利用用户名、头像等属性字段的距离或者频率作为判断依据进行身份识别；基于 UGC 的方法[7,8]利用 UGC 发布地点、时间、写作风格等信息进行识别；而基于用户关系的方法[9~12]通常利用用户的邻域特征计算待识别节点之间的跨网络相似度。

为了进一步提升算法的准确率，众多学者开始尝试将上述三类信息进行融合。Kong 等人[13]将该问题形式化定义为锚链接(anchor links)预测问题，综合了 UGC 的文本相似度、时空信息和网络结构，然后训练二分类器进行实现。Zhang 等人[14]提取用户属性特征和网络结构特征，充分考虑局部一致性和全局一致性构建了基于能量的异质网络用户身份识别模型。Liu 等人[15]利用了用户属性、UGC 以及社交行为等各种可用资源提出了一种半监督多目标统一框架。

融合多种信息的算法一定程度上提升了准确率，然而过度严苛的匹配条件导致算法的召回率不高，同时加重了计算的负担。文献[16]指出，用户在网络中的社交行为，如关注行为或信息发布行为都真实地体现了用户的兴趣倾向。尽管用户会根据不同的目的参与网络，但他们的社交行为是由行为习惯和特性驱使的，兴趣倾向是他们潜在意识的表现，在不同的网络中会保持相对稳定。而用户不同的个性使得用户兴趣成为用户之间相互区分的有效信息。由此，挖掘并分析用户兴趣对于识别不同网络中属于同一用户的账号有着非常

**收稿日期：**2018-08-14；**修回日期：**2018-10-13

**作者简介：**邓诗琦（1994-），女，重庆人，硕士研究生，主要研究方向为社交网络分析、数据挖掘（shiqiDeng@163.com）；李雷（1976-），男，河南南阳人，讲师，主要研究方向为智能信息处理、数据挖掘；施化吉（1964-），男，浙江台州人，教授，主要研究方向为社交网络分析、数据挖掘、信息检索、数据库应用技术、信息系统、电子商务.

重要的意义。

为此，本文提出了基于用户兴趣的跨网络用户身份识别算法（UI-UI），旨在依据用户兴趣实现跨网络用户身份识别。该算法首先利用分块思想将待识别的账号划分到不同的数据块中，只考虑同一数据块内账号的匹配可能性，减少盲目匹配次数，降低时间复杂度；其次根据用户 UGC 主题倾向和交互倾向对用户兴趣进行建模，并计算待匹配账号之间的兴趣相似度；最后利用 UI-UI 算法进行账号匹配。实验表明，本文所提算法综合性能显著提高，验证了算法的有效性。

## 1　问题描述与相关定义

### 1.1　跨网络用户身份识别

**定义 1**　用户身份。用户身份是现实世界中可以相互区分的真实个体在网络中拥有的账号。

**定义 2**　跨网络用户。在两个社交网络中均拥有账号的用户称为跨网络用户。

源网络和目标网络分别是两个完整的、存在一定跨网络用户的社交网络。两者的具体定义如下：

**定义 3**　源网络和目标网络。将源网络和目标网络抽象为有向图，分别用 $G_s(V_s,E_s)$ 和 $G_t(V_t,E_t)$ 表示，其中 $V_s$ 和 $V_t$ 分别为两者的账号节点集合，$E_s$ 和 $E_t$ 分别为两者的节点间转发关系集合。

如源网络内部的节点 $u$ 指向节点 $v$ 表示 $u$ 转发了 $v$ 的消息，即 $u$ 对 $v$ 发布的消息感兴趣。

**定义 4**　跨网络用户身份识别。跨网络用户身份识别指从目标网络中挖掘出与源网络中的节点属于同一跨网络用户的节点，即寻找两个网络中属于同一跨网络用户的节点匹配对。

假如存在一个节点匹配对 $(v,w)$，说明源网络 $G_s$ 中的节点 $v$ 与目标网络 $G_t$ 中的节点 $w$ 是匹配的，即 $v$ 和 $w$ 代表的两个账号属于同一现实用户。

### 1.2　基于先验节点的跨网络用户身份识别算法

**定义 5**　先验节点。先验节点（priori nodes， PN）是网络中身份已被识别的节点。

利用身份已知的先验节点识别网络中身份未知节点的方法叫做基于先验节点的跨网络用户身份识别算法。此类算法的突出优点是时间复杂度明显低于无先验节点的算法。

本文将源网络和目标网络的先验节点集分别表示为 $PN_s$ 和 $PN_t$。

## 2　基于用户兴趣的跨网络用户身份识别算法 UI-UI

### 2.1　分块处理

现有算法在识别跨网络用户身份时大多需要对两个网络中的全部节点进行两两匹配，计算量很大且难以扩展到大规模社交网络中。针对这一缺点，UI-UI 算法在进行身份识别之前先对节点进行初筛选，以降低时间开销并增强算法的可扩展性。

实体识别领域中为避免对整个数据集进行笛卡尔集级别的计算提出了分块技术[17,18]。其通过代价较小的预处理，将可能匹配的数据对象分配到同一个数据块中，不可能匹配的数据对象分配到不同的块中，只进行块内数据对象之间的比较，从而降低时间复杂度。

借鉴这种思想，UI-UI 算法首先利用稀疏性较低、获取较容易的用户属性信息对网络中的节点进行分块处理，将源网络和目标网络中用户属性相似度大于阈值的节点划分在一个数据块内。其中各属性相似度的计算方法如下:

a)用户名相似度。

用户名是最易获取也是利用价值很高的属性信息，已有学者[19,20]研究了仅利用用户名识别跨网络用户身份的方法，并取得了较好的效果。

本文采用 Jaro-Winkler 相似度[21]计算用户名相似度。Jaro-Winkler 是一种字符串匹配算法，对于计算短字符串相似度非常有效，计算结果为 1 表示两个字符串完全匹配，结果为 0 则两者没有相似性。则两个字符串 $s_1$ 和 $s_2$ 的 Jaro-Winkler 距离计算如式（1）所示。

$$d_{j-w}=\begin{cases}0, m=0\\ \frac{1}{3}\left(\frac{m}{|s_1|}+\frac{m}{|s_2|}+\frac{m-\frac{t}{2}}{m}\right), m>0\end{cases} \tag{1}$$

其中：$|s_1|$ 和 $|s_2|$ 分别是两字符串的长度；$m$ 是两者匹配的字符数；$t$ 是换位的字符数目。

b)性别、地址等属性。

对于性别、地址等属性值有限且固定的属性，本文采用精确匹配的方法，若两个属性值完全匹配则相似度为 1，否则为 0。

c)空属性值。

用户的隐私设置可能导致很多属性信息不可取，本文为这些空属性值添加一个缺省标记。为缺省的属性值添加特殊标记“none”，并设定其与其他属性值的相似度为 1，表示该缺省属性值有可能与任何属性值相同。

### 2.2　用户兴趣建模

有研究表明，用户在社交网络上会对某些特定主题的内容表现出更强的关注，不同的用户对同一主题的偏好程度不同，且同一用户的兴趣倾向在不同网络中的表现是相似的。因此用户兴趣的差异可以作为用户之间区分的标志。

一方面，用户会在社交网络中发表体现自己兴趣的言论或参与感兴趣内容的讨论，因此合理利用 UGC 中的信息能够有效建模用户兴趣；另一方面，用户会选择与兴趣相投的用户进行交互，因此用户的社交行为也能体现用户兴趣。针对上述分析，本文分别利用 UGC 和用户社交行为对用户兴趣建模，并由此定义兴趣相似度作为跨网络用户身份识别的依据。

#### 2.2.1 主题兴趣建模

UGC 是用户发表在社交网络上的原创内容，文献[16]指出语言特征已被证明能够体现人格的差异，即用户兴趣倾向的不同；文献[22]也证明了用户在网络中发表言论的语言特征可以用来识别其在不同网络中的账号。

为了获取用户的主题兴趣，本文先利用潜在狄利克雷分配(Latent Dirichlet allocation, LDA)模型[23]对 UGC 中的文本信息进行建模，得到 z 个虚拟主题，以及用户在各虚拟主题下的概率分布 $T$，再基于 JS 散度[24]计算用户之间的主题兴趣相似度。

**定义 6**　主题兴趣相似度。若主题个数为 z，且节点 $v\in G_s$ 和 $w\in G_t$ 的主题概率分布分别表示为 $T_v=\{t_v^1, t_v^2,\cdots, t_v^z\}$ 和 $T_w=\{t_w^1, t_w^2,\cdots, t_w^z\}$，则两者的主题兴趣相似度如式（2）所示。

$$\begin{aligned} sim_{\mathrm{top}}(v,w)&=1-D_{JS}(T_v,T_w)\\ &=1-\frac{1}{2}\left[D_{KL}(T_v,\frac{T_v+T_w}{2})+D_{KL}(T_w,\frac{T_v+T_w}{2})\right]\end{aligned} \tag{2}$$

其中：$D_{KL}(T_v,T_w)=\sum_{l=1}^{z}t_v^l\log\frac{t_v^l}{t_w^1}$ 为两个分布的 KL 散度；且有 $sim_{\mathrm{top}}(v,w)\in[0,1]$，该值越大，则节点 $v$ 和 $w$ 的主题兴趣相似度越高。

2.2.2 交互兴趣建模

根据同质理论[25]，用户会选择与其“相似”的其他用户进行交互。基于这种交互机制社交网络为用户提供了关注和转发功能，这也使得用户的社交行为成为挖掘用户兴趣的一种有效途径。现有算法[9,13]认为两个账号的关注列表越相似，两者属于同一用户的可能性越高。然而用户的关注行为具有一定的时效性和任意性，即用户的兴趣随时间而变化，且用户关注的目的可能不是因为兴趣，而是出于互惠性[26]或其他目的。因此本文采用更能体现用户交互兴趣的转发关系[27]进行建模，并利用式（3）计算节点间的交互兴趣相似度。

**定义 7**　交互兴趣相似度。基于 Jaccard 系数[28]，节点 $v \in G_s$ 和 $w \in G_t$ 的交互兴趣相似度定义如下：

$$sim_{rel}(v,w)=\frac{|(I_v \cap PN_s)\cap(I_w \cap PN_t)|}{|(I_v \cap PN_s)\cup(I_w \cap PN_t)|} \tag{3}$$

其中：$I_v$ 是节点 $v$ 转发过的消息所属的节点集；同理 $I_w$ 为节点 $w$ 转发过的消息所属的节点集；$PN_s$ 和 $PN_t$ 分别为源网络和目标网络的先验节点集。交互兴趣相似度越大两个节点各自交互的用户越相似，两者属于同一跨网络用户的概率越高。

2.2.3 兴趣相似度计算

根据前文对主题兴趣相似度与交互兴趣相似度的计算，如下定义节点兴趣相似度:

$$sim_{int}(u,v)=\alpha sim_{top}(u,v)+(1-\alpha)sim_{rel}(u,v) \tag{4}$$

由于社交网络中用户的主题倾向是较为稳定的，而在算法刚开始时可利用的先验节点数量较少，对于用户交互信息的获取不利，所以在此以黄金分割比例设置调和因子 $\alpha$=0.618。

## 2.3　UI-UI 算法

UI-UI 算法首先根据用户属性信息进行分块处理，得到相应的数据块，目标网络中与源网络节点 $v$ 同属一个数据块的节点的集合即为节点 $v$ 的候选节点集；然后计算节点 $v$ 与其候选节点集中各节点的兴趣相似度，选取相似度最高的节点 $w$ 作为节点 $v$ 的匹配节点，将选出的目标网络中的节点 $w$ 作为待识别的节点，同样根据兴趣相似度在源网络中寻找匹配，若匹配到的节点为 $v$，则匹配成功，将匹配对 $(v,w)$ 输出，并将 $v$ 和 $w$ 分别加入先验节点集 $PN_s$ 和 $PN_t$；最后采取迭代的思想不断更新节点匹配对集，直至无新匹配对生成，算法结束。具体算法流程如算法 1 所示。

算法 1　基于用户兴趣的跨网络用户身份识别算法 UI-UI

输入：源网络 $G_s(V_s,E_s)$、目标网络 $G_t(V_t,E_t)$、源网络先验节点集 $PN_s$、目标网络先验节点集 $PN_t$。

输出：节点匹配对集 $P$。

a)利用用户属性信息对两个网络中的节点进行分块处理。

b)对于源网络 $G_s$ 中的待识别节点 $v \in V_s - PN_s$，根据分块结果确定其候选节点集，并计算其与候选节点集各节点之间的兴趣相似度 $sim_{int}$。

c)选择与节点 $v$ 兴趣相似度最高的节点 $w$ 作为 $v$ 的待匹配节点。

d)将目标网络看做源网络进行反向验证，即将目标网络中待匹配节点 $w$ 作为待识别节点，在源网络中寻找匹配。

(a)若节点 $w$ 匹配到节点 $v$，则认为匹配成功，并将 $v$ 加入 $PN_s$，$w$ 加入 $PN_t$，$(v,w)$ 加入 $P$；

(b)若节点 $w$ 未匹配到节点 $v$，视为匹配失败，返回步骤 b)，继续识别其他待识别节点。

e)循环迭代至无新的节点匹配对生成，输出集合节点匹配对集 $P$。

## 2.4　算法时间复杂度分析

假定源网络 $G_s(V_s,E_s)$ 和目标网络 $G_t(V_t,E_t)$ 中节点集的大小分别为 $|V_s|=m$ 和 $|V_t|=n$，且两个网络中先验节点集大小分别为 $|PN_s|=p$ 和 $|PN_t|=q$，显然有 $p \ll m, q \ll n$，此时两个网络中待识别节点的个数分别为 $a=m-p$ 和 $b=n-q$。

UI-UI 算法时间复杂度的计算分为分块处理、主题建模及节点匹配三个部分。分块处理过程的时间复杂度为 $O(mn)$；主题建模过程的时间复杂度为 $O(m+n)$；而在进行节点匹配时，需要计算每个待识别节点与其候选节点集中全部节点的兴趣相似度。假设 $B_s^i$ 和 $B_t^i$ 分别为源网络和目标网络的一组对应的数据块，其中 $i \in [1,k]$，$i,k \in N^*$，且 $|B_s^1|+|B_s^2|+\ldots+|B_s^k|=a$，$|B_t^1|+|B_t^2|+\ldots+|B_t^k|=b$。这样一来每轮迭代只需计算每个待识别节点与其对应数据块中节点的兴趣相似度，假设最大的一对数据块大小分别为 $|B_s^i|=max\{|B_s^1|,\ldots,|B_s^k|\}=r \ll a$ 和 $|B_t^i|=max\{|B_t^1|,\ldots,|B_t^k|\}=l \ll b$，则计算兴趣相似度的时间复杂度为 $O(rl)$。同理可得反向认证阶段的时间复杂度也为 $O(rl)$)。因此 UI-UI 算法在节点匹配阶段的时间复杂度为 $O(rl)$。

综合上诉分析可知，UI-UI 算法的总时间复杂度为 $O(mn)$。

# 3　实验

## 3.1　实验数据

本文将国外流行社交网站 Facebook 和 Twitter 作为实验对象以验证算法性能。为获取两个网络中真实的用户身份信息，本文首先将提供了用户在各个社交网站上的个人主页链接的 Google+网站作为收集信息的基站，从中收集了 56 107 个用户账号，并从其用户属性信息中提取其在 Facebook 和 Twitter 中的非空且有效的主页链接；然后再分别从 Facebook 和 Twitter 中收集对应用户的属性信息、前 200 条（不足 200 取全部）推文信息和转发信息。获取的数据集具体情况如表 1 所示。

表 1　Facebook 和 Twitter 网络数据

Table 1　Facebook and Twitter network data

| 网络 | 节点数 | 转发连边数 | 跨网络用户数 |
|---|---|---|---|
| Facebook | 5 649 | 12 997 | 1 193 |
| Twitter | 8 373 | 44 376 | |

## 3.2　实验评价标准

本文采用准确率 precision、召回率 recall、综合指标 F1 以及运行时间 running time 作为算法性能的衡量标准。其中前三个指标（简称为 PRF 值）的相关定义如下：

$$precision=\frac{tp}{tp+fp}$$

$$recall=\frac{tp}{tp+fn}$$

$$F1=\frac{2*precision*recall}{precision+recall}$$

其中：$tp$ 是指算法识别出的正确账号个数；$fp$ 是指算法识别错误账号个数；$fn$ 是指算法未识别出的正确账号个数。

## 3.3　实验结果与分析

为保证实验结果的可靠性，避免偶然性，本文实验数据均为算法在相应条件下重复 10 次后的平均值。同时文中实验均假设只存在一对一匹配，即一个用户在一个网络中至多只能拥有一个账号。

3.3.1 先验节点的影响

由于 Facebook 和 Twitter 网络都较为稀疏，选取度数很

小的节点作为先验节点会产生冷启动问题，所以本文只选择入度不小于 $\rho$ 的节点作为先验节点，并设置先验节点的选取比例从 2%~10%分别进行实验。图 1 展示了 $\rho$=50 时算法性能随先验节点比例变化的影响；图 2 展示了先验节点比例为 8%时，$\rho$ 变化对算法性能的影响。

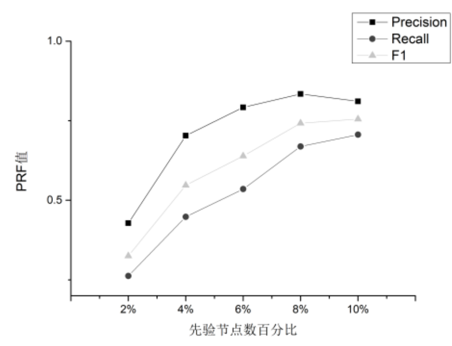



图 1　$\rho$=50 时，PRF 值随先验节点数的变化

Fig. 1　PRF changes with number of PN when $\rho$=50

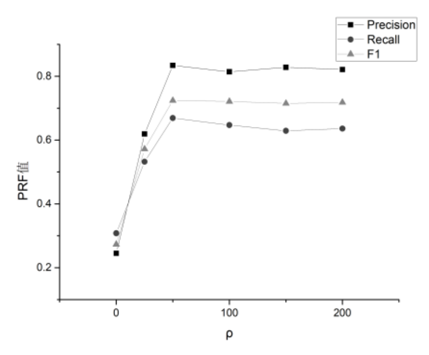



图 2　先验节点比例为 8%时，PRF 值随 $\rho$ 的变化

Fig. 2　PRF changes with $\rho$ when proportion of PN is 8%

据图 1 所示，当先验节点比例为 2%时，算法的准确率和召回率都很低，F1 值仅为 0.325。这是由于当先验节点较少时，可用来识别节点的信息不 v 增加节点兴趣的辨识度提升，各项指标都明显上升。同时可以注意到算法的准确率在先验节点比例为 8%时达到峰值 0.834，而此后继续增加先验节点数算法的准确率略有下降，原因是当先验节点过多时，反而会导致与待识别节点兴趣相似度一样的节点变多，这些节点难以区分对匹配结果造成干扰。

实验结果表明，当先验节点比例大于等于 8%时 UI-UI 算法均能取得 0.81 以上的准确率以及 0.67 以上的召回率。

如图 2 所示，$\rho$=0 时算法的准确率和召回率仅分别为 0.245 和 0.308，因为不限制先验节点的最低度数会导致一部分度数很低的节点被选中，这些低度数的节点不利于节点交互兴趣的获取，一方面使得节点交互兴趣辨识度不高，识别准确率低；另一方面产生冷启动问题使得算法召回率受到很大影响。而随着 $\rho$ 值的增加算法性能明显提升，且算法的各项指标在 $\rho$=50 后趋于稳定。

实验结果证明了设定先验节点最低度数的有效性，当设置 $\rho \geq 50$ 时，UI-UI 算法能达到 0.81 以上的准确率以及 0.62 以上的召回率。

### 3.3.2 算法性能对比

为了进一步验证算法性能，本文将提出的 UI-UI 算法与两个现有识别效果较好的算法进行性能比较。第一种对比算法是 FRUI 算法[10]，其利用了共有已知好友数量实现跨网络节点的一对一匹配；第二种对比算法是 MNA 算法[13]，该算法提取了 UGC 的时空信息以及文本相似度，并训练了 SVM 进行身份识别。

实验设置 UI-UI 算法中 $\rho$=50，先验节点比例为 8%；FRUI 算法中先验节点比例也设置为 8%；而 MNA 算法是监督学习算法无须设置先验节点。实验结果如图 3 所示。

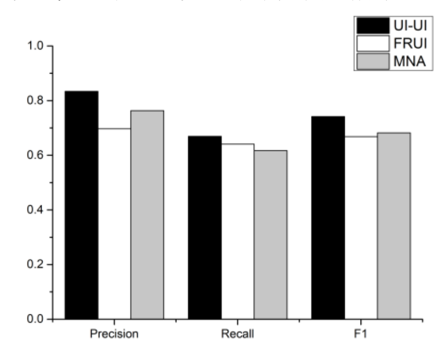



图 3　PRF 值对比

Fig. 3　PRF comparison

如图 3 所示，UI-UI 算法的准确率和召回率都优于另两种算法，且综合评价指标 F1 值比 FRUI 算法提升了 11.1%，比 MNA 算法提升了 8.8%。

FRUI 算法准确率明显低于另两种算法，这是因为该算法仅利用用户的好友关系进行身份识别，无法区分网络中大量结构相似的节点，而 UI-UI 和 MNA 算法都融合了 UGC 信息和用户交互信息，大大增加了节点的辨识度，准确率显著提升。但是 MNA 算法利用 UGC 的时空信息进行识别，这类信息在社交网络中非常稀疏，虽然相比纯拓扑算法提升了一定准确率，但算法召回率很低，也难以扩展到大规模的社交网络中。相比 FRUI 和 MNA 这两种对比算法而言，UI-UI 算法摈弃了传统利用好友关系挖掘用户交互信息的方式，而采用了更能体现用户兴趣偏好的转发关系进行设计，同时利用主题建模方法挖掘用户隐藏在 UGC 信息中的兴趣偏好，准确率和召回率都有明显提升。据图 3 所示，UI-UI 算法的识别准确率超出 FRUI 算法 19.7%，超出 MNA 算法 9.3%；召回率超出 FRUI 算法 4.4%，超出 MNA 算法 8.4%。实验结果证明了用户兴趣对于识别跨网络用户身份的有效性。

本文用总运行时间来评估算法的效率。如图 4 所示，相同数据集上的实验结果表明 UI-UI 算法的运行时间比 FRUI 算法略高，但明显低于 MNA 算法，少于 MNA 算法所需时间的一半。这是由于 FRUI 算法仅考虑了网络拓扑结构，而 UI-UI 算法融入了 UGC 信息，主题建模的过程增加了一定的时间开销。但对比同样融合多种网络信息的 MNA 算法，UI-UI 算法中分块预处理为算法的匹配过程节约了大量时间。

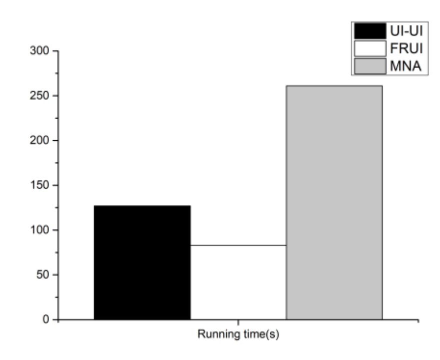



图 4　运行时间对比

Fig. 4　Running time comparison

综上所述，UI-UI 算法相比现有的跨网络身份识别算法准确率和召回率更优，且时间开销更少，更适用于大规模的社交网络。

## 4 结束语

现有算法没有考虑到人类本质上的兴趣偏好对于跨网络身份识别的有效性，为此本文利用用户发布的文本信息以及其转发关系中隐含的兴趣倾向对用户兴趣进行建模，以此定义节点跨网络的相似性；此外，算法在节点匹配之前加入了节点初筛选阶段，通过用户属性对于节点进行分块处理，减少了大量匹配计算，降低了运行时间，使算法更适用于大规模社交网络。实验结果表明，本文所提算法在综合性能和运行时间上均具有明显的优势，验证了用户兴趣是识别跨网络用户身份的一种有效特征。当然，本文仍有许多值得进一步研究的地方，如用户兴趣的建模方式还有很多，如何更加精确有效地提取用户兴趣；用户的兴趣是有一定时效性的，如何将时间因素融入到用户兴趣建模中等。

## 参考文献：


[1] Carmagnola F, Cena F. User identification for cross-system personalisation [J]. Information Sciences, 2009, 179 (1): 16-32.

[2] Deng Zhengyu, Sang Jitao, Xu Changsheng. Personalized video recommendation based on cross-platform user modeling [C]// Proc of IEEE International Conference on Multimedia and Expo. Piscataway, NJ: IEEE Press, 2013: 1-6.

[3] Yan Ming, Sang Jitao, Mei Tao, *et al.* Friend transfer: cold-start friend recommendation with cross-platform transfer learning of social knowledge [C]// Proc of IEEE International Conference on Multimedia and Expo. Piscataway, NJ: IEEE Press, 2013: 1-6.

[4] Kumar S, Zafarani R, Liu Huan. Understanding user migration patterns in social media [C]// Proc of AAAI Conference on Artificial Intelligence. [S. l. ] : AAAI Press, 2011: 1204-1209.

[5] Zhang Haochen, Kan Min-Yen, Liu Yiqun, *et al.* Online social network profile linkage [C]// Proc of Asia Information Retrieval Symposium. Berlin: Springer, 2014: 197-208.

[6] Mu Xin, Zhu Feida, Wang Jianzong, *et al.* User identity linkage by latent user space modelling [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l. ] : ACM Press, 2016: 1775-1784.

[7] Riederer C, Kim Y, Chaintreau A, *et al.* Linking users across domains with location data: theory and validation [C]// Proc of International Conference on World Wide Web; International World Wide Web Conferences Steering Committee. 2016: 707-719.

[8] Goga O, Lei Howard, Parthasarathi S H K, *et al.* Exploiting innocuous activity for correlating users across sites [J]. 2013: 447-458.

[9] Man Tong, Shen Huawei, Liu Shenghua, *et al.* Predict anchor links across social networks via an embedding approach [C]// Proc of International Joint Conference on Artificial Intelligence. [S. l. ] : AAAI Press, 2016: 1823-1829.

[10] Zhou Xiaoping, Liang Xun, Zhang Haiyan, *et al.* Cross-platform identification of anonymous identical users in multiple social media networks [J]. IEEE Trans on Knowledge & Data Engineering, 2016, 28 (2): 411-424.

[11] 吴铮，于洪涛，黄瑞阳，等. 基于隐藏标签节点挖掘的跨网络用户身份识别 [J]. 计算机应用研究，2018, 35 (4): 1191-1196. (Wu Zheng, Yu Hongtao, Huang Ruiyang, *et al.* User identification across multiple networks based on hidden label nodes mining [J]. Application Research of Computers, 2018, 35 (4): 1191-1196. )

[12] Feng Shuo, Wang Qian, Shen Derong, *et al.* User identification across social networks based on global view features [C]// Proc of the 14th Web Information Systems and Applications Conference. Piscataway, NJ: IEEE Press, 2017: 2018: 93-98.

[13] Kong Xiangnan, Zhang Jiawei, Yu Philip S. Inferring anchor links across multiple heterogeneous social networks [C]// Proc of ACM International Conference on Information & Knowledge Management. [S. l. ] : ACM Press, 2013: 179-188.

[14] Zhang Yutao, Tang Jie, Yang Zhilin, *et al.* COSNET: connecting heterogeneous social networks with local and global consistency [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l. ] : ACM Press, 2016: 1485-1494.

[15] Liu Siyuan, Wang Shuhui, Zhu Feida, *et al.* HYDRA: large-scale social identity linkage via heterogeneous behavior modeling [C]// Proc of ACM SIGMOD International Conference on Management of Data. [S. l. ] : ACM Press, 2014: 51-62.

[16] 张磊，陈贞翔，杨波. 社交网络用户的人格分析与预测 [J]. 计算机学报, 2014, 37 (8): 1877-1894. (Zhang Lei, Chen Zhenxiang, Yang Bo. Personality analysis and prediction of social network users [J]. Chinese Journal of Computers, 2014, 37 (8): 1877-1894. )

[17] Kenig B, Gal A. Efficient entity resolution with MFIBlocks [Z]. 2009.

[18] Cvhristen P. A Survey of Indexing technique s for scalable record linkage and deduplication [J]. IEEE Trans on Knowledge & Data Engineering, 2012, 24 (9): 1537-1555.

[19] Zafarani R, Liu Huan. Connecting users across social media sites: a behavioral-modeling approach [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l. ] : ACM Press, 2013: 41-49.

[20] 刘东，吴泉源，韩伟红，等. 基于用户名特征的用户身份同一性判定方法 [J]. 计算机学报，2015, 38 (10): 2028-2040. (Liu Dong, Wu Quanyuan, Han Weihong. *et al.* User identification across multiple websites based on username features [J]. Chinese Journal of Computers, 2015, 38 (10): 2028-2040. )

[21] Jaro M A. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida [J]. Publications of the American Statistical Association, 1989, 84 (406): 414-420.

[22] Almishari M, Tsudik G. Exploring linkability of user Reviews [C]// Lecture Notes in Computer Science. 2012: 307-324.

[23] Blei D M, Ng A Y, Jordan M I. Latent dirichlet al location [J]. Journal of Machine Learning Research Archive, 2003, 3: 993-1022.

[24] Eissa T, Razak S A, Ngadi M D A. Towards providing a new lightweight authentication and encryption scheme for MANET [J]. Wireless Networks, 2011, 17 (4): 833-842.

[25] Mcpherson M, Smith-Lovin L, Cook J M. Birds of a feather: homophily in social networks [J]. Annual Review of Sociology, 2001, 27 (1): 415-444.

[26] Kumar R, Novak J, Tomkins A. Structure and evolution of online social network [C]// Proc of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Datamining. New York: ACM Press, 2006: 611-617.

[27] Webberley W, Allen S, Whitaker R. Retweeting: a study of message-forwarding in twitter [C]// Proc of Workshop on Mobile and Online Social Networks. Piscataway, NJ: IEEE Press, 2011: 13-18.

[28] Jaccard P. The distribution of the flora in the alpine zone 1 [J]. New Phytologist, 1912, 11 (2): 37-50.